Electronic Version
Stylesheet Version v1.1.1

Claims

[c1] 1. A method for computing a measure of similarity between a first (or input) document and a second (or search results) document, comprising:
(a) receiving a first list of rated keywords extracted from the first document and a second list of rated keywords extracted from the second document;
(b) using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list;
(c) computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second document;
(d) using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage.

[c2] 2. The method according to claim 1, wherein the second

percentage at (c) is computed by giving weight only to those keywords and their set of neighboring keywords in the first list that match in the second list and a threshold percentage of the keywords in their set of neighboring keywords.

[c3] 3. The method according to claim 2, wherein the second percentage at (c) is computed by giving full weight to those keywords in the first list of rated keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords.

[c4] 4. The method according to claim 2, wherein the threshold percentage is reduced when the first list of rated keywords is identified using OCR.

[c5] 5. The method according to claim 1, further comprising (e) if the first computed percentage does not indicate that the first document is included in the second document, computing a third percentage using the Jaccard distance measure.

[c6] 6. The method according to claim 5, further comprising (f) if the third computed percentage indicates that the first document is a revision of the second document, computing a fourth percentage indicating what percentage of keyword ratings along with a set of their neigh-

boring keyword ratings in the second list also exist in the first list.

[c7] 7. The method according to claim 6, further comprising using the fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords.

[c8] 8. The method according to claim 1, wherein the first computed percentage indicates that the first document is included in the second document when the percentage defined by ratio of Sum1/Sum2 is greater than approximately ninety percent, where:
D1 is the number of keywords in first list of keywords;
D2 is the number of keywords in the second list of keywords;
Sum1 is the sum of the weights of keywords that appear in D1 that also appear in D2;
Sum2 is the sum of the weights of keywords in D1.

[c9] 9. The method according to claim 1, wherein the first list of rated keywords includes one or more keywords trans-

lated from a second language different from a first language that is identified as being a primary language of the first document.

[c10] 10. The method according to claim 1, wherein the first document is a portion of the second document.

[c11] 11. A system for computing a measure of similarity between a first (or input) document and a second (or search results) document, comprising:
(a) means for receiving a first list of rated keywords extracted from the first document and a second list of rated keywords extracted from the second document;
(b) means for using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list;
(c) means for computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second document;
(d) means for using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed per-

centage.

[c12] 12. The system according to claim 11, wherein the second percentage at (c) is computed by said computing means by giving weight only to those keywords and their set of neighboring keywords in the first list that match in the second list and a threshold percentage of the keywords in their set of neighboring keywords.

[c13] 13. The system according to claim 12, wherein the second percentage at (c) is computed by said computing means by giving full weight to those keywords in the first list of rated keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords.

[c14] 14. The system according to claim 12, wherein the threshold percentage is reduced when the first list of rated keywords is identified using OCR.

[c15] 15. The system according to claim 11, further comprising (e) if the first computed percentage does not indicate that the first document is included in the second document, means computes a third percentage using the Jaccard distance measure.

[c16] 16. The system according to claim 15, further comprising (f) if the third computed percentage indicates that

the first document is a revision of the second document, means computes a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list.

[c17] 17. The system according to claim 16, further comprising means for using the fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords.

[c18] 18. The system according to claim 11, wherein the first computed percentage indicates that the first document is included in the second document when the percentage defined by ratio of Sum1/Sum2 is greater than approximately ninety percent, where:
D1 is the number of keywords in first list of keywords;
D2 is the number of keywords in the second list of keywords;
Sum1 is the sum of the weights of keywords that appear in D1 that also appear in D2;
Sum2 is the sum of the weights of keywords in D1.

[c19] 19. The system according to claim 11, wherein the first list of rated keywords includes one or more keywords translated from a second language different from a first language that is identified as being a primary language of the first document.

[c20] 20. An article of manufacture for computing a measure of similarity between a first (or input) document and a second (or search results) document, the article of manufacture comprising computer usable media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:

(a) receiving a first list of rated keywords extracted from the first document and a second list of rated keywords extracted from the second document;

(b) using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list;

(c) computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second

document;

(d) using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage.